УДК 81.139

Атрибутирование русскоязычных текстов с использованием закона больших чисел

Филимонов В. В.[1*], Амиева А. М.[1], Живодёров А. А.[2,3], Крамаренко А. А.[2]

[1]Уральский федеральный университет, каф. полиграфии и веб-дизайна, ул. Мира, 32, Р041, Екатеринбург, Россия, 620002
[2]Уральский федеральный университет, каф. технической физики, Мира, 31, Екатеринбург, Россия, 620078
[3]Центральная научная библиотека УРО РАН, Софьи Ковалевской, 22, Екатеринбург, Россия, 620137

Аннотация. Работа посвящена статистическому исследованию русскоязычных текстов. Исследуется связь между длиной текста и среднеквадратичным отклонением величины χ^2 для распределения сочетаний гласных букв. Установлена справедливость закона больших чисел для этой величины. Работа выполнена на кафедре полиграфии и веб-дизайна ИРИТ-РтФ УрФУ.

Ключевые слова: корпус, статистика χ^2, стандартное отклонение, длина текста, кластер.

Attribution is a Russian-language texts using the law of large numbers

Filimonov V. V.[1*], Amieva A. M.[1], ZhivodyorovA. A.[2,3], Kramarenko A. A.[2]

[1]Ural Federal University, Mira, 32, R041, Ekaterinburg, Russia, 620002
[2]Ural Federal University, Department of Technical physics, Mira, 31, Ekaterinburg, Russia, 620078
[3]Central scientific library UB RAS, Sofia Kovalevskaya, 22, Ekaterinburg, Russia, 620137

Abstract. The work is devoted to the statistical study of Russian texts. Examines the relationship between text length and standard deviation of the χ^2 values for the distribution of combinations of vowels. Established the validity of the law of large numbers for this value. The work was performed at the Department of printing and web design IRIT-RTF UrFU.

Keywords: case, χ^2 statistics, the standard deviation, the length of the text, cluster.

Введение

В работе [1] в целом подтвердилась выдвинутая нами гипотеза о том, что существуют структурные особенности текста, которые позволяют без учёта смыслового содержания отнести текст к определённому жанру. Эти особенности выражены в величине статистики $\chi 2$, которая связана с частотными характеристиками появления комплексов гласных букв.

Исследования проводились на материале из специально созданного «Корпуса текстов русского языка» (далее Корпус), который включает в себя на сегодняшний день около 1 300 текстов художественного, научного, социально-политического, административного, религиозного направлений, а также из газетных и журнальных публикаций. Каждое направление представлено в виде соответствующего подкорпуса. Для переводных текстов указано двойное авторство: автор первоначального текста и автор перевода [2].

Результатом работы стало распределение текстов Корпуса по нескольким интервалам, связанным с величиной $\chi 2$ и прагматикой самих текстов. Оказалось возможным достаточно отчётливо выделить несколько кластеров, которые условно могут быть названы: поэзия, художественная проза, научные, социально-политические и административные тексты.

При отнесении текста к тому или иному кластеру экспертная оценка совпала с машинной. Это несмотря на то, что машина не ориентируется ни на смысл текста, ни на его название и анализирует только последовательность знаков. Таким образом, экспериментально подтвердилась адекватность метода статистики $\chi 2$ для анализа русскоязычных текстов.

Однако границы между кластерами оказались нечёткими, диффузными, т.е. существуют области, в которых присутствуют тексты, принадлежащие различным жанрам. По-видимому, кластеризация текстов по одному параметру не во всех случаях позволяет однозначно отнести текст к определённому жанру, в некоторых случаях можно говорить лишь о вероятной атрибуции текста.

Различия между величинами статистики $\chi 2$ могут носить случайный характер и быть связанными с конечностью длины текста. Чем больше длина текста, тем меньше должно быть стандартное отклонение (SD) значений $\chi 2$ и в случае «достаточно больших» текстов оно асимптотически стремится к нулю. Под «достаточно большим» мы понимаем текст, длина которого существенно больше средней длины текстов, представленных в Корпусе. Средняя длина текстов по Корпусу составляет 133 081 гласных.

Согласно закону больших чисел

$$SD = \frac{c}{\sqrt{N}}, \qquad (1)$$

где N — размер выборки (в нашем случае — количество гласных букв в тексте), c — коэффициент пропорциональности.

Определив стандартное отклонение значений $\chi 2$ для текстов различной длины, мы сможем построить доверительные интервалы (95 %) для $\chi 2$

$$\Delta\chi^2 = \chi^2 \pm \frac{1{,}96c}{\sqrt{N}}, \quad (2)$$

а также ответить на вопрос: случайным или закономерным является различие этих значений для каждой пары текстов, т.е. вероятнее всего сможем уточнить границы кластеров.

1. Ход работы

Задачей настоящей работы является определение зависимости стандартного отклонения значений $\chi 2$ от длины текста, т.е. определение величины коэффициента c в формуле (1).

Из формулы (1) видно, что коэффициент c не зависит от длины текста N и может быть связан с особенностями самого текста или быть характерным для языка в целом.

Для исследования были выбраны по несколько текстов из разных подкорпусов Корпуса (таблица 3) и два сгенерированных при помощи программы «Rondo» 1. Оба сгенерированных текста состоят из одних гласных букв. Один из них (текст 1) был сгенерирован, исходя из равновероятного их появления ($\omega = 1/9 = 0{,}111\ldots$), другой (текст 2) — исходя из средних частот появления гласных букв в русскоязычных текстах. Значения частот взяты из работы [1] и представлены в таблице 1. Сгенерированные тексты выступают в роли образцов, которые не принадлежат никакому стилю, и с которыми можно будет сравнивать реально существующие тексты.

Таблица 1. Средние частоты появления отдельных гласных букв в тексте

Буква	Средняя частота (ω)
а	0,183
е	0,206
и	0,167
о	0,259
у	0,064
ы	0,046
э	0,008
ю	0,016
я	0,051

[1]Программа «*Rondo*» была специально написана для исследования сотрудником ЦНБ УрО РАН Л.Г. Горбичем

Длины текстов представлены в таблице 2.

Таблица 2. Длины сгенерированных текстов

Текст	Длина текста
Текст 1	899 828
Текст 2	817 555

Из каждого текста были взяты фрагменты по 100 000, 200 000, 300 000, 400 000, 500 000, 600 000, 700 000, 800 000 гласных букв, по 9 фрагментов для каждого количества гласных.

Для каждой такой группы рассчитывалась величина $\chi 2$ для троек гласных букв (букворазмещение по три). Под «букворазмещением по три» мы понимаем три последовательно появляющиеся в тексте гласные буквы. Первое букворазмещение составляют первая, вторая и третья гласные буквы в тексте, второе — вторая, третья и четвёртая, третье — третья, четвёртая, пятая и т. д.

Значение $\chi 2$ рассчитывается по формуле

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i^{\text{theor}} - n_i^{\text{emp}}\right)^2}{n_i^{\text{theor}}} \cdot \frac{50000}{N}, \tag{3}$$

где n_i^{theor} представляет собой количество троек, рассчитанных из соображений независимого появления отдельных букв, n_i^{emp} было получено в результате пересчёта троек в реальном тексте, $k = 729$ (т. е. 9^3) — общее количество вариантов троек букв, N — длина текста (количество гласных в тексте), нормировка на 50 000 гласных применяется для того, чтобы значения $\chi 2$ для текстов разной длины можно было сравнивать между собой.

Далее для каждого текста были посчитаны стандартные отклонения величины $\chi 2$

$$S = \sqrt{\frac{\sum (x - M)^2}{(n - 1)}}, \tag{4}$$

где x — значение признака у каждого объекта в группе, M — средняя арифметическая признака, n — число вариант выборки [3].

Согласно закону больших чисел, величина стандартного отклонения $\chi 2$ при увеличении количества гласных букв в тексте будет асимптотически стремиться к нулю, согласно

$$S = \frac{c}{\sqrt{N}}, \tag{5}$$

где S — величина стандартного отклонения, N — количество гласных в тексте, c — эмпирическая константа.

Если константа с в различных текстах будет постоянным значением, то её можно связать с особенностями языка в целом. Если же значения с будут разными для разных текстов, то её можно рассматривать как атрибут конкретного текста.

Были найдены зависимости стандартного отклонения величины $\chi 2$ от длины текста и при помощи метода наименьших квадратов были построены соответствующие аппроксимирующие кривые (рис. 1, 2, 3).

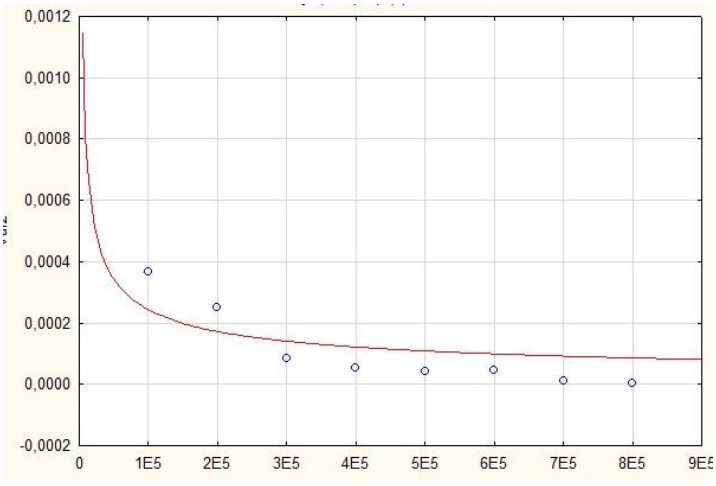

Рис. 1. Текст 1, с = 0,96009

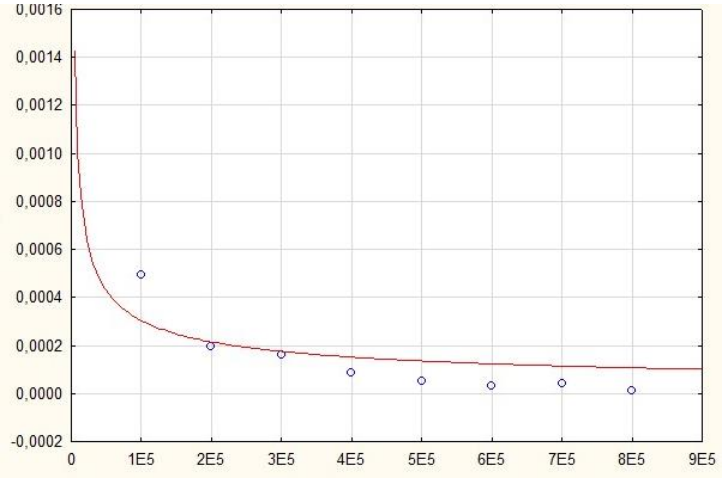

Рис. 2. Текст 2, с = 0,077

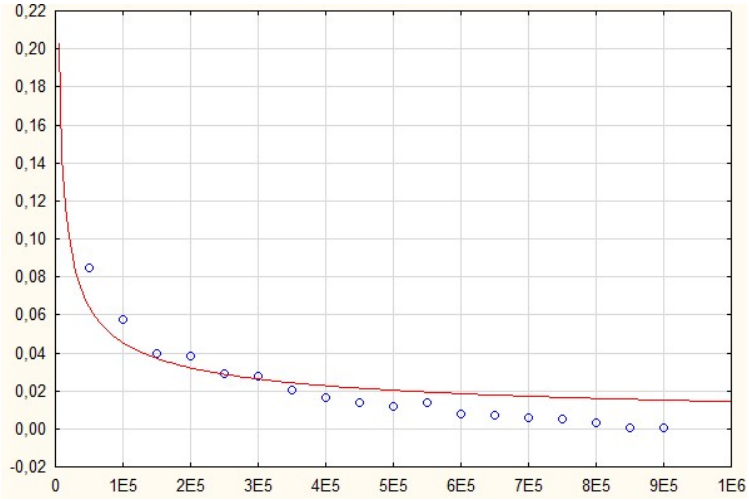

Рис.3. Текст 3, c = 14,3689

Для равновероятностного текста $c = 0{,}96009 \approx 1$, что соответствует закону больших чисел.

Из графиков (рис. 1, 2, 3) видно, что значение с для реальных текстов сильно отличается от соответствующих значений для случайных текстов, это позволяет говорить о том, что коэффициент с является атрибутом конкретного текста.

Таким же образом были исследованы тексты различных направлений: проза, поэзия, научные, религиозные и административные тексты. Полученные значения коэффициента с представлены в таблице 3.

Таблица 3. Значения коэффициента *c*

	Название	Количество гласных	χ^2 по частотам	Коэффициент *c*
Проза				
1	С. Кинг. Рассказы	862 111	0,072776	0,411966
2	Л.Н. Толстой. Анна Каренина	574 339	0,083033	0,901109
3	Жан-Жак Руссо. Юлия, или новая Элоиза, пер. Худядова	568 106	0,082794	1,08854
4	А. Дюма. Граф Монте-Кристо	788 741	0,074963	1,29893
5	А. Дюма. Белые и синие	505 999	0,063976	1,31381
6	А. Азимов, Р.М. Аллен. Калибан, пер. Шестаков	662 428	0,089077	1,44827

	Название	Количество гласных	χ^2 по частотам	Коэффициент *c*
7	В. Шукшин. Рассказы	528 303	0,059566	1,77124
Проза				
8	Д.Ф. Штраус. Жизнь Иисуса	600 153	0,082545	2,09935
9	Н. Чернышевский. Что делать	737 785	0,073976	2,95488
10	М.А. Шолохов. Тихий Дон	965 666	0,072776	14,3689
Поэзия				
1	А.С. Пушкин. Полное собраний стихотворений	195 909	0,040008	0,604469
2	М.И. Цветаева. Полное собраний стихотворений	202 428	0,057582	3,73283
3	М.Ю. Лермонтов. Полное собраний стихотворений	152 838	0,112489	4,97561
Научные				
1	И. Гарин. Век Джойса	634 207	0,087055	1,93798
2	В.М. Блейхер. Толковый словарь психиатрических терминов	601 602	0,244967	2,67757
3	С. Головин. Словарь психолога	891 234	0,205944	3,04607
4	В. Губин. Учебник по философии	706 353	0,147888	3,1944
5	А.Г. Спиркин. Философия: Учебник	605 259	0,142964	3,34832
6	А. Реймон. Введение в философию истории	549 995	0,172062	3,38643
7	З.М. Черниловский. Всеобщая история государства и права	535 491	0,131944	3,64434
8	О. Ермишин. Афоризмы	938 862	0,087021	3,84053
9	История средних веков, под ред. Н.Ф. Колесницкого	510 722	0,147316	3,8821
10	Л.Я. Аверьянов. Социология: искусство задавать вопросы	680 595	0,193382	4,77789
11	Ландау 2,3 том	717 752	0,266705	7,72917

	Название	Количество гласных	χ^2 по частотам	Коэффициент *c*
12	А.Н. Коновалов, Л.Б. Лихтерман, А.А. Потапов. Клиническое руководство по черепно-мозговой травме (том 1)	667 663	0,244477	10,7205
13	К. Маркс. Капитал	957 347	0,224623	20,4576
Религиозные				
1	Жития святых — декабрь	507 922	0,079116	0,314479
2	Иоанн Златоуст. Том 5	522 927	0,098361	0,398052
3	Жития святых — ноябрь	613 458	0,084013	0,958659
4	Иоанн Златоуст. Том 7	754 135	0,098173	1,34662
5	Иоанн Златоуст. Том 9	721 660	0,108005	1,37833
6	Иоанн Златоуст. Том 10	632 256	0,100657	1,42364
7	Иоанн Златоуст. Том 8	506 842	0,117734	1,51282
8	Жития святых — январь	546 184	0,094662	1,82146
9	Иоанн Златоуст. Том 12	807 014	0,095526	1,85055
10	Д. Андреев. Роза мира	557 106	0,110032	1,87649
11	Иоанн Златоуст. Том 3	861 189	0,108176	2,05322
12	Библия	885 879	0,066320	2,55298
13	Иоанн Златоуст. Том 11	738 837	0,103384	2,91758
Административные				
1	Земля и право	334 233	0,293959	5,08531
2	Гражданский кодекс	367 057	0,273041	5,5717
3	Уголовный процесс	430 406	0,281776	5,69543
4	Трудовой кодекс	237 047	0,730052	10,7771
5	Инструкция о порядке предоставления эфирного времени	425 692	0,445479	16,6537
6	О несостоятельности (банкротстве)	557 627	0,558997	38,3672

2. Результаты

1. Подтвердилась справедливость предположения, что среднеквадратичное отклонение для величины $\chi 2$ обратно пропорционально квадратному корню из длины текста, т.е. подчиняется закону больших чисел.

2. Из полученных данных видно, что значение коэффициента с для сгенерированного текста близко к 1 ($c = 0{,}96009$). Для реальных текстов значение с в большинстве случаев лежит в диапазоне от 1,5 до 5, резко выделяются тексты религиозного подкорпуса (с от 0,3 до 3) и административного подкорпуса (с от 5 до 38). Также следует отметить, что тексты по истории и философии имеют близкие значения коэффициентов с (от 3,2 до 3,9), в то время как значения $\chi 2$ для этих текстов сильно различаются, хотя и принадлежат одному (научному) кластеру (от 0,08 до 0,17).

В дальнейшем планируется построение доверительных интервалов для $\chi 2$, что позволит уточнить ранее полученные результаты (уточнить границы кластеров).

Список литературы

1. Филимонов В. В., Амиева А. М., Сергеев А. П. Кластеризация русскоязычных текстов с применением статистики χ^2 // Материалы международной научно-практической конференции (Екатеринбург 12–13 января 2016 г.) Информация: передача, обработка, восприятие. Екатеринбург : УрФУ имени первого Президента России Б.Н. Ельцина, 2016. С. 164–174.
2. Инструменты корпусной лингвистики / Амиева А.М., Филимонов В. В., Сергеев А. П., Тарасов Д. А. // Материалы международной научно-практической конференции (Екатеринбург 14–15 декабря 2015 г.) Информационные технологии, телекоммуникации и системы управления. Екатеринбург : УрФУ имени первого Президента России Б.Н. Ельцина, 2016. С. 251–260.
3. Стандартное отклонение [Электронный ресурс]. URL: http://studopedia.ru/13_130873_standartnoe-otklonenie.html. (Дата обращения: 1.11.2016).